SegCorr: a statistical procedure for the detection of genomic regions of correlated expression.

Eleni Ioanna Delatola [1,2,3,4], Emilie Lebarbier [1,2], Tristan Mary-Huard [1,2,5], François Radvanyi[3,4], Stéphane Robin[1,2], and Jennifer Wong [3,4]

[1]AgroParisTech UMR518, Paris 5e, France
[2]INRA UMR518, Paris 5e, France
[3]Institut Curie, Centre de Recherche, Paris, F-75248 France
[4]CNRS UMR144, Equipe Oncologie Moléculaire, Paris, F-75248 France
[5]INRA, UMR 0320 / UMR 8120 Génétique Végétale et Évolution Le Moulon, F-91190 Gif-sur-Yvette, France

Abstract

Motivation: Detecting local correlations in expression between neighbor genes along the genome has proved to be an effective strategy to identify possible causes of transcriptional deregulation in cancer. It has been successfully used to illustrate the role of mechanisms such as copy number variation (CNV) or epigenetic alterations as factors that may significantly alter expression in large chromosomic regions (gene silencing or gene activation).

Results: The identification of correlated regions requires segmenting the gene expression correlation matrix into regions of homogeneously correlated genes and assessing whether the observed local correlation is significantly higher than the background chromosomal correlation. A unified statistical framework is proposed to achieve these two tasks, where optimal segmentation is efficiently performed using dynamic programming algorithm, and detection of highly correlated regions is then achieved using an exact test procedure. We also propose a simple and efficient procedure to correct the expression signal for mechanisms already known to impact expression correlation. The performance and robustness of the proposed procedure, called SegCorr, are evaluated on simulated data. The procedure is illustrated on cancer data, where the signal is corrected for correlations possibly caused by copy number variation. The correction permitted the detection of regions with high correlations linked to DNA methylation.

Availability and implementation: R package SegCorr is available on the CRAN.
Contact: eldelatola@yahoo.gr

1

1 Introduction

In the last decade, the study of local co-expression of neighbor genes along the chromosome has become a question of major importance in cancer biology (Clark, 2007). The development of "Omics" technologies have permitted the identification of several mechanisms inducing local gene regulation, that may be due to a common transcription factor (De and Babu, 2010) or common epigenetic marks (Stransky et al., 2006; Frigola et al., 2006). Copy number variation due to polymorphism or to genomic instability in cancer is also a possible cause for observing a correlation between neighbor genes (Aldred et al., 2005), as their expressions are likely to be affected by the same copy number variation (CNV). It has further been observed that local regulations may occur in specific nuclear domains, as the nuclear region is an environment which may favor or not transcription (Bickmore, 2013).

Investigating the impact of a specific source of regulation (TF, DNA methylation, histoine modifications, CNV) on the expression has now become a common practice for which statistical tools are readily available. On the other hand, only a few methods have been proposed to focus on the direct analysis of gene expression correlation along the chromosomes. The direct analysis of correlations may have different purposes:

(i) one can aim at detecting all potential chromosomal domains of co-expression, then investigating to which extend known causal mechanism are responsible for the observed co-expression patterns,

(ii) one can aim at detecting chromosomal domains of co-expression where correlations are not caused by already known sources of regulation, in order to identify new potential mechanisms impacting transcription.

Addressing problems (i) and (ii) is crucial to fully understand transcriptional deregulation and/or to model gene regulation.

We first consider problem (i) and provide a precise definition of our purpose: one aims at identifying correlated regions, i.e. blocks of neighbor genes, the expression of which display correlations across patients that are significantly higher than expected. Indeed, it has been observed that background correlation between adjacent genes along the genome does exist. This should not be confounded with the co-expression that can be locally observed due to the aforementioned mechanisms. Note that this definition does not include detection of regions with biased gene expression as studied in Nilsson et al. (2008), Xiao et al. (2009) or Seifert et al. (2014), that account for potential correlations between adjacent genes but do not aim at detecting highly correlated regions.

Several approaches have already been proposed to tackle problem (i). Clustering methods accounting for the chromosomal organization of the genes have been proposed in CluGene (Dottorini et al., 2013) and DIGMAP (Yi et al., 2005). Sliding windows procedures have also been considered, such as G-NEST (Lemay et al., 2012), REEF (Coppe et al., 2006) or TCM (Reyal et al., 2005). The principle is then to compute correlation scores for genes falling within the window, then to detect local peaks of high correlation scores. While these procedures have been successfully applied to cancer data, all tackle the detection of correlated region using heuristics. As such, they

suffer from classical limitations associated with these techniques, including local optimum (for clustering algorithms) or detection instability according to the choice of the window size (for sliding windows).

It is now well known that the problem of finding regions in a spatially ordered signal can be cast as a segmentation problem, for which standard statistical models exist, along with efficient algorithms to find the globally optimal solution. According to our definition, the detection of correlated regions boils down to the block-diagonal segmentation of the correlation matrix between gene expressions. Such an approach has been proposed in image processing (Lu et al., 2013), finance (Lavielle and Teyssière, 2006) and bioinformatics for CNV analysis (Zhang et al., 2010), but to the best of our knowledge it has never been considered for the detection of correlated expression regions.

While problem (i) can be addressed on the basis of only expression data, problem (ii) requires the additional measurement of the signal one needs to account for. For example, consider that one seeks for locally expressed co-regulation events that are not due to copy number variations to include copy number changes due to polymorphism or genomic alteration observed in cancer. The strategy we adopt here consists in first correcting the expression data for potential CNV regulation, then in applying the procedure described to solve problem (i) on the corrected signal. The corrected signal is obtained by regressing the initial expression signal on the CNV signal. Although quite simple, the strategy turns out to be efficient in practice. An alternative strategy would be to jointly model both the expression and the signals to correct for, and then propose within this framework a correction. Such a strategy would necessitate to adapt the modeling to the specific combination of signals one has at hand. In comparison, the regression procedure proposed here can be applied to any kind and any number of signals one needs to correct for.

The outline of the present article is the following. In Section 2 we propose a parametric statistical framework for the problem of correlated region identification. Finding regions of co-regulated genes can then be achieved by maximum likelihood inference (to find the boundaries of each region along with their correlation levels). An exact test procedure to assess the significance of the correlation with respect to background correlation is proposed in Section 3. We introduce a simple procedure to correct expression data beforehand for some known (and quantified) source of correlation. Because the background correlation level is a priori unknown, an estimator of this quantity is also proposed. The performance of the SegCorr procedure is illustrated in Section 4 on simulated data, along with a comparison with the TCM algorithm proposed in Reyal et al. (2005). Finally, a case study on cancer data is presented in Section 5.

3

2 Correlation matrix segmentation

2.1 Statistical model

We consider the following expression matrix:

$$Y = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ Y_{21} & \cdots & Y_{2p} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix}$$

where Y_{ij} stands for the expression of gene j ($j = 1, \ldots, p$) observed in patient i ($i = 1, \ldots, n$). The i-th row of this matrix is denoted Y_i and corresponds to the expression vector of all genes in patient i. In order to detect regions of correlated expression, we consider the following statistical model. Profiles $\{Y_i\}_{1 \leq i \leq n}$ are supposed to be i.i.d, normalized (centered and standardized), following a Gaussian distribution with block-diagonal correlation matrix G:

$$G = \begin{bmatrix} \Sigma_1 & & & & \\ & \ddots & & \mathbf{0} & \\ & & \Sigma_k & & \\ & \mathbf{0} & & \ddots & \\ & & & & \Sigma_K \end{bmatrix}$$

$$\text{with } \Sigma_k = \begin{bmatrix} 1 & \cdots & \rho_k \\ \vdots & \ddots & \vdots \\ \rho_k & \cdots & 1 \end{bmatrix}. \tag{1}$$

The model states that genes are spread into K contiguous regions, with respective lengths p_k ($k = 1, \ldots, K$, $\sum_{1 \leq k \leq K} p_k = p$). Genes belonging to different regions are supposed to be independent, whereas genes belonging to a same region are supposed to share the same pairwise correlation coefficient ρ_k. This amounts to assume that some specific effect (e.g. methylation) affects the expression of all genes belonging to the region. More specifically, let U_k denote the vector of the region effect (accross patients). For all genes j from region k, the model can be written as $Y_{ij} = \mu_{ij} + U_{ik} + E_{ij}$. The error terms E_{ij} are all independent and independent from U_{ik} such that $\mathbb{V}(U_{ik})/\mathbb{V}(Y_{ij}) = \rho_k$.

2.2 Accounting for known sources of regulation

As mentioned in the Introduction, a second task (ii) can be to detect correlated regions which are not due to an already known mechanism. While in the model of the previous section, Y_{ij} denoted the expression of gene j for patient i. In this case, Y_{ij} refers to the gene expression corrected for some underlying mechanism. This correction step can be done using the following regression model :

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \tag{2}$$

where x_{ij} stands for the covariate observed in patient i for gene j. For instance, in the illustration of Section 5.4, x_{ij} is the copy number associated to patient i at location of gene j. The corrected signal is then $\widetilde{Y}_{ij} = Y_{ij} - \widehat{\beta}_0 - \widehat{\beta}_1 x_{ij}$. Note that it suffices to assume that (ϵ_{ij}) are independent among patients (but not among genes) to get the standard linear regression estimates (see Anderson (1958), Chapter 8).

Several articles have investigated the relationship between gene expression and mechanisms such as CNV or methylation, and proposed sophisticated models for this relationship (Menezes et al., 2009; van Wieringen et al., 2010; Leday et al., 2013). Alternatively, gene expression can be predicted using one of these models. Then, these predictions can replace the explanatory variable x in Equation (2). The residuals of the resulting model can be considered as the corrected signal.

2.3 Inference of correlated regions

Parameter inference in Model (1) amounts to estimate the number of regions K, the region boundaries $0 = \tau_0 < \tau_1 < \cdots < \tau_{K+1} = p$, and the correlation parameters $\rho_1, ..., \rho_K$ within each of these regions. Here, we consider a maximum penalized likelihood approach. First, we show that for a given K the optimal region boundaries and correlation coefficients can be efficiently obtained using dynamic programming. The number of regions can then be selected using a penalized likelihood criterion.
For a fixed K, the estimation problem can be formulated as follows:

$$\arg \max_{\tau_1 < \cdots < \tau_K} \max_{\rho_1, ..., \rho_K} \mathcal{L} \tag{3}$$

where the log-likelihood $\mathcal{L}$ is expressed as

$$-2\mathcal{L} \quad = \quad n \log |G| + \text{tr} \left[Y G^{-1} (Y)^\top \right] ..$$

Here, thanks to the block diagonal structure of the correlation matrix in Model (1), the log-likelihood can be rewritten as

$$-2\mathcal{L} \quad = \quad \sum_{k=1}^{K} \left\{ n \log |\Sigma_k| + \text{tr} \left[Y^{(k)} \Sigma_k^{-1} (Y^{(k)})^\top \right] \right\}$$

where $Y^{(k)}$ stands for the set of expression from Y corresponding to genes included in the k-th region. Moreover defining $\mathcal{L}_k$, the log-likelihood in region k containing genes from $\tau_{k-1} + 1$ to τ_k as

$$- 2\mathcal{L}_k \quad = \quad -2\mathcal{L}(\tau_{k-1} + 1, \tau_k) \tag{4}$$

$$= \quad n \log |\Sigma_k| + \text{tr} \left[Y^{(k)} \Sigma_k^{-1} (Y^{(k)})^\top \right] , \tag{5}$$

the optimization problem (3) boils down to

$$\arg \max_{\tau_1 < \cdots < \tau_K} \sum_{k=1}^{K} \max_{\rho_k} \mathcal{L}_k. \tag{6}$$

5

Inference when K is fixed   We first show that for a given region k with known boundaries, explicit expressions can be obtained for both the ML estimator $\widehat{\rho}_k$ and the likelihood $\mathcal{L}_k$ at the optimum:

Lemma 1 *The maximum of $\mathcal{L}_k$ with respect to ρ_k is reached for*

$$\widehat{\rho}_k = \frac{\sum_j^{p_k} \sum_k^{p_k} \widehat{G}_{jk} - p_k}{p_k^2 - p_k} \tag{7}$$

where $\widehat{G}_{jk} := n^{-1} \sum_{i=1}^n Y_{ij} Y_{ik}$. Furthermore, the maximal value of $\mathcal{L}_k$ is given by

$$
\begin{aligned}
-2\widehat{\mathcal{L}}_k &= n \left[p_k + (p_k - 1) \log \left(-\frac{\sum_j^{p_k} \sum_k^{p_k} \widehat{G}_{jk} - p_k^2}{p_k^2 - p_k} \right) \right. \\
&\quad \left. + \log \left(\frac{\sum_j^{p_k} \sum_k^{p_k} \widehat{G}_{jk}}{p_k} \right) \right].
\end{aligned}
\tag{8}
$$

The proof is given in Appendix A. The expression of Problem (6) is now

$$\arg \max_{\tau_1 < \cdots < \tau_K} \sum_{k=1}^K \widehat{\mathcal{L}}_k \tag{9}$$

which is additive with respect to the $\widehat{\mathcal{L}}_k$ terms that can be straightforwardly computed thanks to Lemma 1. Consequently, optimization can be performed via Dynamic Programming (DP, Lavielle (2005), Picard et al. (2005)). The optimal boundaries, and correlation estimators can be obtained at computational cost $\mathcal{O}(Kp^2)$.

Lasso-type approaches have been proposed to tackle segmentation problems in a faster way (see e.g. Tibshirani and Wang (2008)). First, note that such methods rely on a relaxation of the original problem, so that the result may be different from the exact solution of problem (8). Furthermore, as for matrix segmentation, such approaches have been proposed (Bien and Tibshirani (2011); Levina et al. (2008)), which do not allow to capture the longitudinal structure (i.e. blocks of neighbor genes).

Model selection.   To choose the number of regions, we adopt the model selection strategy proposed in Lavielle (2005). For each $1 \le K \le K_{\max}$, we define the maximal log-likelihood for K regions as

$$L_K = \max_{\tau_1 < \cdots < \tau_K} \sum_{k=1}^K \widehat{\mathcal{L}}(\tau_{k-1} + 1, \tau_k).$$

Furthermore, the normalized log-likelihood is defined as

$$\widetilde{L}_K = \frac{L_{K_{\max}} - L_K}{L_{K_{\max}} - L_1} (\widetilde{K}_{\max} - \widetilde{K}_1) + 1,$$

where $\widetilde{K}_j = 5 \times j + 2 \times j \log(p/j)$ is the penalty function. Lavielle (2005) suggests to estimate the number of regions $\widehat{K}$ as the value of K such that $\widetilde{L}_K$ displays the largest slope change. Namely, we take

$$\widehat{K} = \arg\min_K \left\{ (\widetilde{L}_K - \widetilde{L}_{K+1}) - (\widetilde{L}_{K+1} - \widetilde{L}_{K+2}) > S \right\},$$

where the value of threshold S is predefined. Throughout the paper, $S = 0.7$ as suggested in Lavielle (2005). The robustness of the results with respect to other values for threshold S is investigated in Section 4. This global approach (dynamic programming and model selection) has been applied with success for CNV detection (see Picard et al. (2005) and Lai et al. (2005) for a comparative study.)

3   Assessing correlation significance

It has been observed (Cohen et al., 2000; Spellman and Rubin, 2002; Reyal et al., 2005; Stransky et al., 2006) that background correlations may exist between adjacent genes along the genome, i.e. one expects the correlation level in any region to be positive. As a consequence, one has to check whether a given region exhibits a correlation level that is significantly higher than the background correlation level ρ_0, that is observed by default.

Test procedure.   Once the correlation matrix segmentation is performed, it is possible to identify regions with high correlation levels by testing $H_0 : \rho_k = \rho_0$ vs $H_1 : \rho_k > \rho_0$. This can be done using the following test statistic for region k:

$$T_k = \frac{1}{n} \sum_i^n \left(Y_{i\bullet}^{(k)} - Y_{\bullet\bullet}^{(k)} \right)^2$$

where

$$Y_{i\bullet}^{(k)} = \sum_{j=\tau_{k-1}+1}^{\tau_k} Y_{ij} \text{ and } Y_{\bullet\bullet}^{(k)} = \sum_{i=1}^n Y_{i\bullet}^{(k)} .$$

Assuming Model (1) is true, test statistic T_k has distribution

$$T_k \sim \lambda(p_k, \rho_k)\chi_{n-1}^2 \text{ where } \lambda(p_k, \rho_k) = \frac{(1 + (p_k - 1)\rho_k)}{np_k} .$$

Here χ_{n-1}^2 stands for the chi-square distribution with $n - 1$ degrees of freedom. The proof is given in Appendix B. We emphasize that this test is exact and does not rely on any resampling strategy.

Consequently, the p-value associated to region k is given by

$$\mathbb{P}\left(\lambda(p_k, \rho_0)Z > T_k^{obs}\right) \text{ , where } Z \sim \chi_{n-1}^2.$$

7

Statistical power. We now study the ability of the proposed test to detect a region with width p where the correlation ρ is higher than in the background. The probability to detect such a region depends on both p and ρ, and writes

$$
\begin{aligned}
Po(n, p, \rho) &= \Pr\{T > \lambda(p, \rho_0) q_{n-1,1-\alpha}\} \\
&= \Pr\left\{Z > \frac{\lambda(p, \rho_0)}{\lambda(p, \rho)} q_{n-1,1-\alpha}\right\}
\end{aligned}
$$

where $Z \sim \chi_{n-1}^2$ and $q_{n-1,1-\alpha}$ is the $1 - \alpha$ quantile for the χ_{n-1}^2 distribution. Figure 1 (Top) displays the evolution of power for different values of p and ρ. Here ρ_0 and n are fixed at 0.15 and 58, respectively, which correspond to the values observed in the reference dataset (see Section 5). The nominal levels of α are 5%, 0.5% and 0.05%. These correspond to realistic thresholds, once multiple testing corrections such as Bonferroni or FDR are performed. One can observe that even for small values of ρ, the power is high whatever the nominal level as soon as the number of genes in the considered region is equal to or higher than 5. It also shows that the procedure will probably fail to find regions of size 3, if the correlation is not at least as high as 0.7 (to obtain a power of 0.8). On the same graph (Bottom), one observes that a sample of size 50 is sufficient to efficiently detect regions of size 5, as long as the correlation is higher than 0.6. Larger samples will be required if one wants to efficiently detect regions with smaller correlation levels.

Background correlation estimation. The test procedure requires the knowledge of parameter ρ_0 that is unknown in practice. However, it can be estimated using

$$
\widehat{\rho_0} = |\underset{i>1}{\operatorname{median}}(\operatorname{corr}(Y_{i-1}, Y_i))| . \tag{10}
$$

Under the assumption that most pairs of adjacent genes display a ρ_0 correlation, i.e. only a few number of regions with moderate sizes exhibit a high level of correlation, $\widehat{\rho_0}$ is a robust estimator of the background correlation. The behavior of estimator (10) is investigated in Section 4.

4 Simulation study

In this section, we first study the quality of the proposed estimator of ρ_0. Then we study the ability of SegCorr to detect correlated regions and compare its performance with this of TCM algorithm. The robustness of the method with respect to the choice of the model selection threshold S will be investigated in Section 5.2 on real data, since very little difference were observed on the simulated data (results not shown).

4.1 Simulation design

For each round of simulation, a sample of $n = 58$ profiles with 22 chromosomes each is generated, with a number of genes per chromosome identical to the one observed in the reference dataset (see description in Section 5.1). Each chromosome is split into K_{chr} regions, according to the segmentation obtained in Stransky et al. (2006). Each
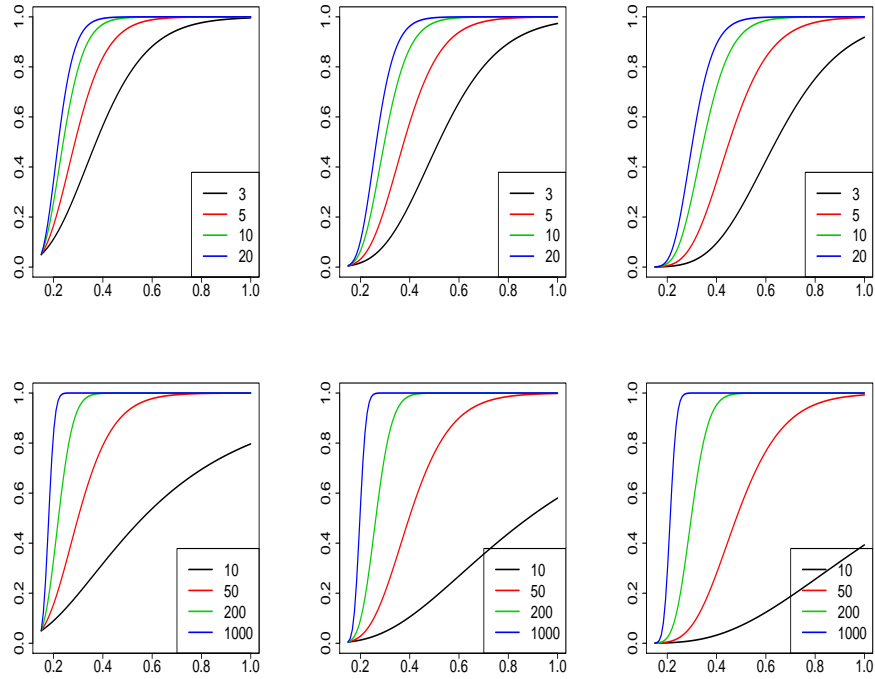
Figure 1: Top: Power curves as a function of ρ, for a fixed cohort size $n = 58$ and varying region width $p = 3, 5, 10, 20$. Bottom: Same graphs for a region of fixed width $p = 5$ but varying cohort sizes $n = 10, 50, 200, 1000$. In all graphs ρ_0 is fixed at 0.15. The nominal level α of the test is set to 5% (left), 0.5% (center), 0.05% (right)
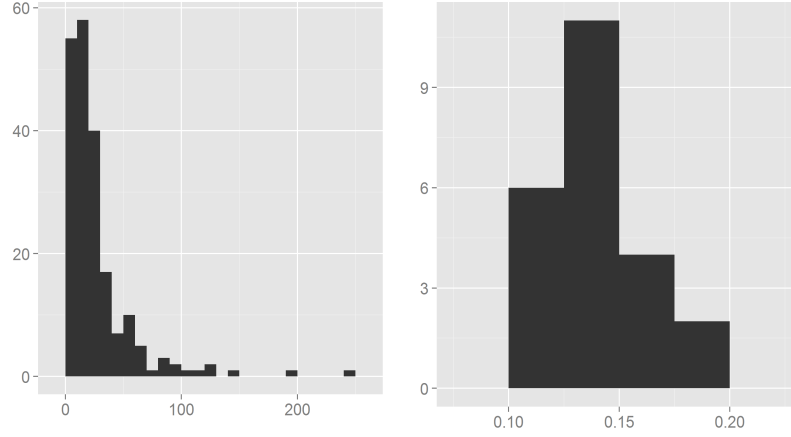
Figure 2: Left: Length of H_1 regions in the reference dataset. Right: Distribution of the backgroung correlation $\hat{\rho}_0$ obtained from the reference data according to the segmentation obtained in Stransky et al. (2006).

segmentation alternates between H_0 regions, i.e. regions with nominal background correlation ρ_0, and H_1 regions with nominal correlation ρ_1. Figure 2 depicts the length of the H_1 correlated regions obtained in the Stransky et al. (2006) study. As it can be seen, this length varies substantially from a H_1 region to another. For ρ_1, different values were considered (ranging from 0.3 to 0.9), all H_1 regions of all chromosomes sharing this same ρ_1 coefficient. For ρ_0, two cases are considered :

- Scenario 1 (Easy case): ρ_0 is the same for all H_0 regions of all chromosomes (3 different values considered: 0.08, 0.18, 0.28).

- Scenario 2 (Realistic case): ρ_0 is identical for all H_0 regions within a chromosome, but varies from a chromosome to another. The specific value chosen for a given chromosome is the one obtained for this same chromosome on the reference dataset. The distribution of ρ_0 across chromosomes on the reference dataset is given in Figure 2.

Additionally, correlations between genes from different regions were observed on the reference dataset. Thus, a similar extra block diagonal correlation pattern was generated in the simulations, with a level of background correlation between blocks similar to the one within blocks.

Lastly, for each combination (ρ_0, ρ_1) the simulation was replicated 20 times.

4.2   Quality of the ρ_0 estimator

For this study, we consider Scenario 1. Figure 3 illustrates the estimation accuracy of ρ_0 under different levels of both H_0 and H_1 correlations on chromosome 3. Estimator (10) yields in over-estimated values of the true background correlation level. One observes
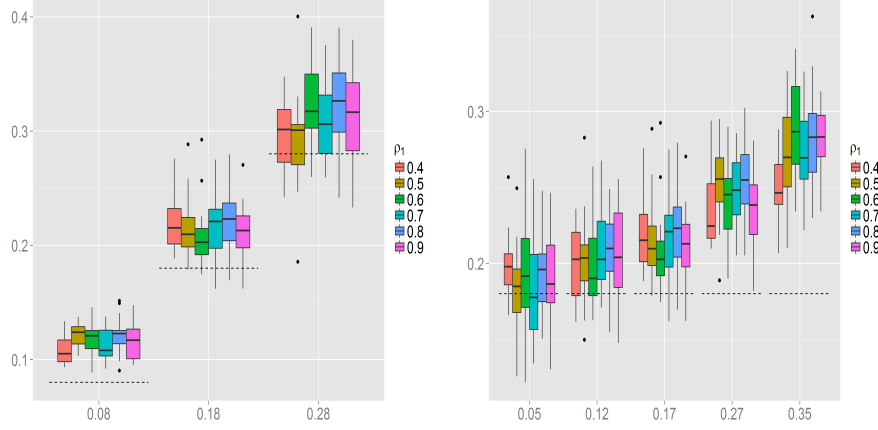
Figure 3: Left: estimation of ρ_0 for chromosome 3 under different levels of both H_0 and H_1 correlations ($\rho_0 = 0.08, 0.18$ and 0.28). Dashed lines indicate the true ρ_0. Right: estimation of ρ_0 for $\rho_0 = 0.18$ and different levels of H_1 correlations according to the fraction of H_1 correlations (the results are showed for five typical chromosomes only).

that the overestimation does not depend on the correlation level in H_1 regions, thanks to the use of the median. Still, as expected, it is linked to the proportion of pairs of adjacent genes with H_1 correlations, as showed in Figure 3. Importantly, while overestimation of ρ_0 will result in a decrease of power, it will not increase the false positive rate (FDR or FWER).

4.3   Performance evaluation

To assess the performance of SegCorr, the true positive rate (TPR = sensitivity), false positive rate (FPR = 1− specificity) and area under the ROC curve (AUC) were considered. These criteria were first computed at the gene level. However, as the goal is to identify correlated regions, a definition of TPR and FPR at the region level was adopted. We considered the intersection between the true and the estimated segmentations and computed the number of true/false positive/negative regions. This amounts at classifying each gene into one of four status (true/false × positive/negative) and then to merge neighbor genes sharing a same status into regions. The status of a region is given by the status of its genes. Consequently, criteria computed at the region level are more stringent as they measure the precision of region boundary estimation.

Figure 4 shows the AUC for the first simulation scenario under various configurations, with ρ_1 fixed at 0.5. When ρ_0 is between 0.08 and 0.18, most regions are correctly detected. For $\rho_0 = 0.28$ (a value higher than what is observed on the reference dataset, see Figure 2), the task becomes difficult and the performances deteriorate.

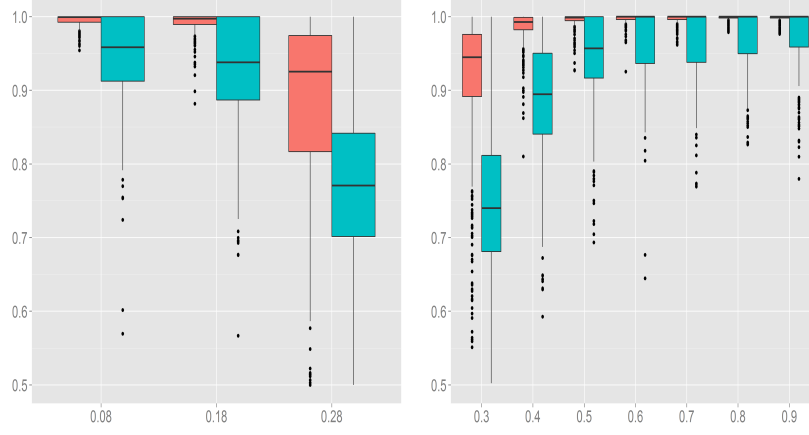On the second simulation study, the behavior of SegCorr was explored under dif-

Figure 4: AUC at the gene level (red) and region level (blue). The higher the AUC the better. Left: Simulation design 1 with fixed $\rho_1 = .5$ (x-axis: ρ_0). Right: Simulation design 2 (x-axis: ρ_1).

ferent ρ_1. Obviously the task becomes easier when ρ_1 gets larger. Figure 4 shows that SegCorr performs well when $0.5 \leq \rho_1 \leq 0.9$. When $\rho_1 \leq 0.5$, (remind that the background correlation can be as high as 0.2, see Figure 2) although the performances remain good at the gene level, the boundaries of the regions are detected less accurately.

4.4 Comparison with the TCM algorithm

SegCorr was compared with the TCM algorithm introduced by Reyal et al. (2005) for the detection of local correlations. In the literature, many methods tackling the same problem as the TCM have been proposed. The choice of the TCM as a competing method was based on the availability of the code. Figure 5 displays the AUC achieved by SegCorr and TCM under Scenario 1. When ρ_0 is large ($\rho_0 = 0.28$), one observes that the mean performance of both methods are comparable with higher variability for SegCorr at the gene level and at the region level for TCM. Since the aim is to detect regions rather that genes, the SegCorr procedure seems more appropriate. For small or medium values of background correlations ($\rho_0 = 0.08, 0.18$) SegCorr achieves better AUC than TCM at both the gene and the region levels. As a conclusion, SegCorr appears to be a more consistent and efficient procedure to detect correlated regions.

5 Bladder cancer data

In this section, we apply SegCorr on the dataset described in Section 5.1. It is now well known that copy number variation (CNV) impacts gene expression (Sebat et al., 2004). Here our goal is to detect regions where the correlation is not due to CNV. Therefore we correct the expression signal for CNV variation according to the strategy described
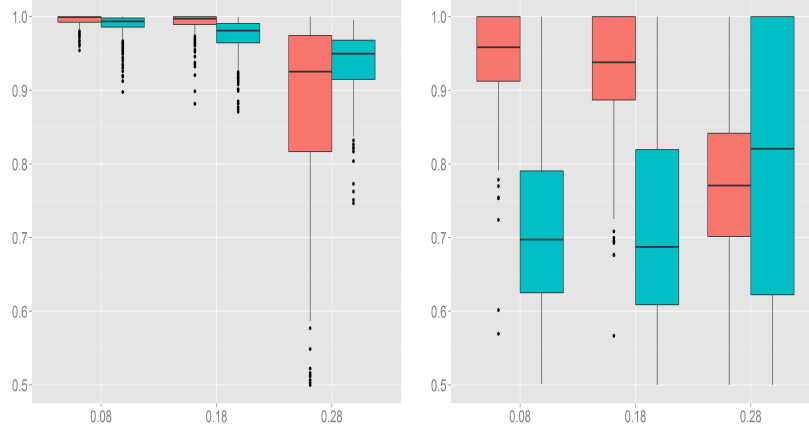
Figure 5: AUC of the SegCorr (red) and TCM (blue) algorithms for the simulation scenario 1 as a function of ρ_0. Left: gene level. Right: region level.

in Sections 2.1 and 5.3. The effect of this correction is investigated in Section 5.4. Lastly, Section 5.5 illustrates the biological results obtained after correction for CNV.

5.1   Data presentation

The dataset consists of $n = 58$ bladder tumors. Gene expression were measured using exon 1.0 Affymetrix arrays and RMA normalisation Irizarry et al. (2003) was applied. The number of genes per chromosome ranges from a 293 to 2192 (with average 950). Additionally genomic and methylation data were collected for the same tumors. Genomic data were obtained with Illumina Human610-Quad SNP arrays and methylation data with Illumina Human methylation 450k arrays. For the latter, the normalization procedure proposed by Teschendorff et al. (2013) was used. The combination of probe positions was made according to the custom BrainArray EntrezGene annotation (Dai et al., 2005).

5.2   Study of the model selection threshold S

For the model selection criterion (see section 2.3), the threshold S must be tuned in such a way to avoid under/over-segmentation. The smaller the value of S the higher the number of segments. As stated in Section 2.3, S was fixed to 0.7 as advocated in Lavielle (2005). Figure 6 shows the evolution of the number and location of H_1 regions detected by SegCorr according to S on a typical chromosome (chromosome 3). One can see that most of these H_1 regions are stable for values of S between 0.6 and 0.9.
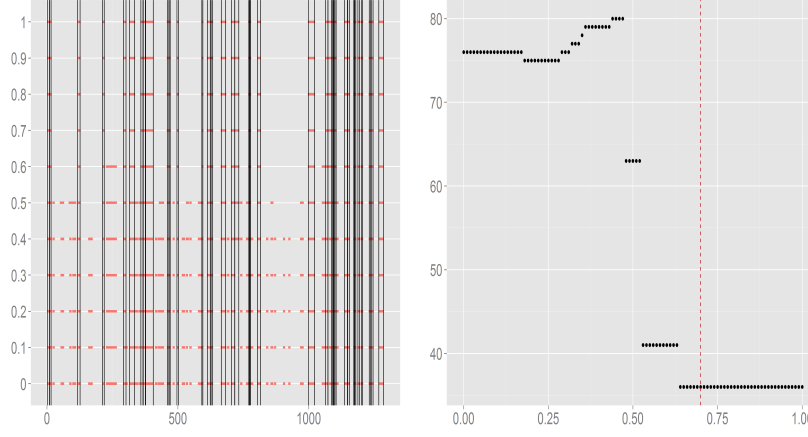
Figure 6: Left: statistically significant regions in red obtained for different values of S. The vertical lines correspond to the ones obtained with the default value of S we considered ($S = 0.7$). Right: number of statistically significant regions for different values of S. The dotted vertical red line corresponds to $S = 0.7$.

5.3 Procedure for CNV correction

To correct the expression signal from CNV, one first needs to detect the CNV regions from the SNP signal. To this aim, we consider the segmentation method proposed by Picard et al. (2011) implemented in the R package cghseg. Denote SNP_{it} the SNP signal of patient i at position t, the model writes

$$SNP_{it} = \mu_{ik} + E_{it} \text{ if } t \in I_k^i = [t_{k-1}^i + 1, t_k^i]. \tag{11}$$

where the E_{it} are i.i.d centered Gaussian with variance σ^2. The method estimates the number of regions, the boundaries of the regions, denoted $\hat{t}_k^i$ and the signal mean within each region k in patient i, denoted $\hat{\mu}_{ik}$.

We then use the regression model (2) to make the correction where x_{ij} is the mean $\hat{\mu}_{ik}$ obtained previously if the SNP position $\hat{t}_k^i$ corresponds to gene j of the expression signal in patient i. Since the SNP and expression signals are not aligned, there might be either one, many or no SNP probes that belong to the corresponding gene region. We then propose to define x_{ij} as follows : if one or many probes are related to gene j, mean $\hat{\mu}_{ik}$ or the average of the different means is considered respectively; if there is no probe, a linear interpolation is performed.

5.4 CNV Dependent Region

We first investigate the effect of CNV correction (described in Section 5.3) by comparing the results obtained on the raw and corrected signals. Figure 7 displays the number of significant H_1 regions as a function of the test level α for both the raw and corrected

14

Chrom.	Genes
2	ASAP2, ITGB1BP1, CPSF3, IAH1, ADAM17, YWHAQ, TAF1B
6	MBOAT1, E2F3, CDKAL1, SOX4, LINC00340
12	MDM2, CPM, CPSF6, LYZ, YEATS4, FRS2, CCT2
16	COG7, GGA2, EARS2, UBFD1, NDUFAB1, PALB2, DCTN5

Table 1: Four examples of CNV-dependent regions

signals. For small values of α (which are typically used for testing significance), the number of detected regions are quite similar. However, only one third of the detected genes are common, meaning that the regions detected with the two signals are quite different. Furthermore, as the correction remove all effects due to CNV, the estimated background correlation is lower in the corrected signal than in the raw signal (mean decrease across all chromosomes of .02). This makes the test we propose more powerful and explains why, while CNV-due regions are removed, the number of detected regions for given α remains about the same.

To illustrate this phenomenon more precisely, we considered a set of four regions in chromosomes 2, 6, 12 and 16 known to be associated with CNV in bladder cancer (Heidenblad et al., 2008; Network et al., 2014). These regions, given in Table 1, are detected by SegCorr when applied to the raw expression data. When considering the corrected signal, these regions are not detected any more. For example, when considering the region in chromosome 2, the background correlation was $\widehat{\rho}_0 = 0.163$ and the correlation within this region was $\widehat{\rho}_k = 0.561$, resulting in a highly significant p-value: 2.1e-7. After correction we get $\widehat{\rho}_0 = 0.132$ and $\widehat{\rho}_k = 0.284$, which results in a non-significant p-value: 2.5e-2.

More generally, over the 184 regions solely detected on the raw signal with p-value smaller than 5% (before multiple testing correction), more than a half (98) get non significant when considering the corrected signal. This explains a substantial part of the difference between the regions detected on raw and corrected signals. This also shows that the proposed CNV correction strategy performs reasonably well.

5.5 CNV-independent regions

General description. When applied to the CNV-corrected expression signal, SegCorr detected 569 significant regions (p-value adjusted ≤ 0.05) which are distributed throughout the genome (an average of 23 regions per chromosome). Among these regions, 158 regions contained well known gene family clusters such as the HOXA, HOXB, HOXD clusters, several KRT clusters, the epidermal differentiation complex, and HLA gene families clusters whose expression is known to be co-regulated (Sproul et al., 2005). We next undertook a Gene Ontology terms analysis and identified an enrichment of genes belongs to the keratinization pathway (p-value 4.09E-19 and FDR q-value 9.01E-16). The expression of this pathways is strongly associated with a subgroup of bladder cancer called basal-like bladder cancer (Rebouissou et al., 2014). 11/28 CNV independent regions detected by TCM using different platforms (Affymetrix U95A for
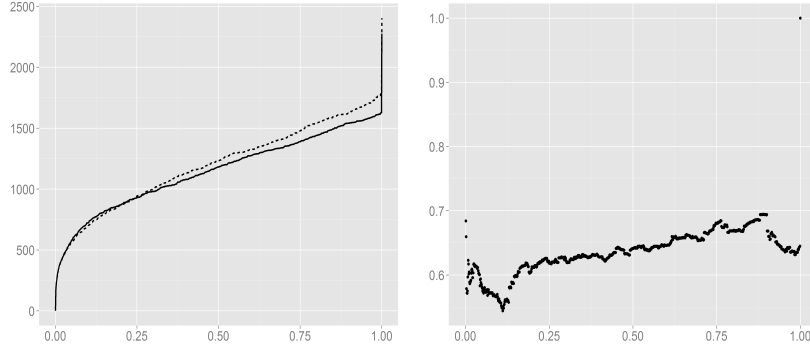
Figure 7: Left: Number of statistically significant regions as a function of α (solid line: corrected signal, dotted line: raw signal). Right: proportion of significant genes common in the two signal as a function of α.

the transcriptome and BAC arrays for the genomic alterations) were also detected by SegCorr (Stransky et al., 2006).

An example of epigenetic region.   We now present a region where the observed correlation is not due to CNV but can be associated with an epigenetic mark. When applied to the CNV corrected expression data, SegCorr detects a region of four genes (HOXB3, HOXB-AS3, HOXB5, HOXB6: $\widehat{\rho}_k = 0.93$, p-value = 3.1e-8) in chromosome 17. This region has already been studied by Vallot et al. (2011) and has been referred to as 17-7. Figure 8 (left) shows a clear pattern in the expression data, which is detected by SegCorr. The right panel provides the DNA methylation data for the same region, which also depicts a clear pattern. This suggest that this region is silenced by an epigenetic mechanism associated with DNA methylation.

This statement is supported by Figure 9 which depicts the correlation between each gene expression and the methylation signal at each locus within the region. It shows a large proportion methylation sites being negatively correlated with the expression of the four genes. This indicates that high DNA methylation level in this region is associated with the silencing of these genes.

6   Discussion

The identification of co-regulated chromosomal regions has many implications in biology. In this paper, we developed a method to identify these regions and we applied it to cancer data. The method relies on a formal definition of what correlated regions are. It takes advantage of an efficient algorithm and a statistical test, which are both exact. We also proposed a correction strategy that allows us to investigate the possible causes of the observed correlations.

Using this method, we could identify copy number dependent and copy number independent correlated regions of expression. Copy number dependent regions correspond
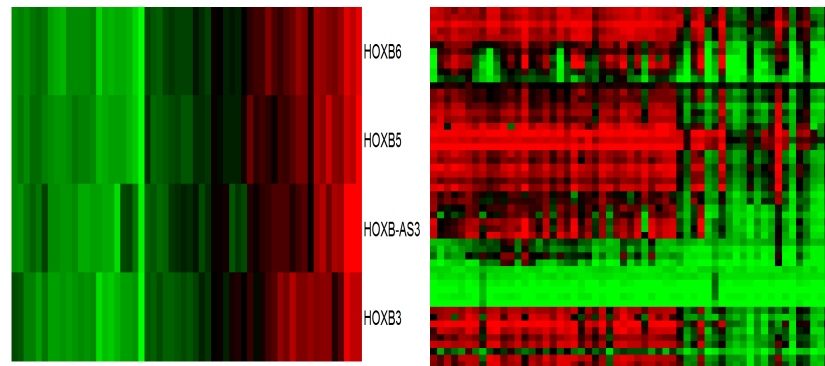
Figure 8: Expression (left) and methylation (right) data from Region 17-7. The ordering of the patients (x-axis) is kept the same in the two plots.
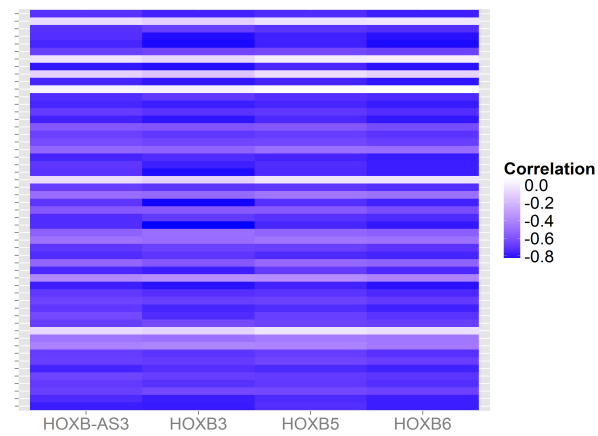

Figure 9: Correlations between expression and methylation data from Region 17-7.

to genomic alterations; copy number independent regions could be due to different mechanisms, including epigenetic mechanism. We showed, for one region, which is part of the HOXB complex, that there is negative correlation between expression and DNA methylation. The detected regions should be further investigated to better understand the underlying mechanism.

While the expression data used here were acquired using the microarray technology, any other technology, including RNA-seq, can be used as well.

In our analysis, we have assumed stretches of correlated contiguous neighboring genes. This is obviously a simplification. Within a correlated region, a gene (or a few genes) could exhibit a weak or even a negative correlation with the other genes. This could occur for different reasons: the gene can be not expressed; alternatively, the gene could be non affected by the regulation process that impacts the other ones; finally, the gene could be impacted in a opposite way compared with the other ones. Note that genes that exhibit no expression or no variation in the dataset can be detected and could be discarded before applying the analysis. While this preprocessing was not required in the present study, running the analysis without removing non-expressed genes would lower the performance of any method aimed at finding correlated (and reasonably homogeneous) regions. Alternatively, looking for the effect of adding a variable number of uncorrelated genes in correlated regions is an obvious follow-up of the present work.

The proposed correction strategy could easily be generalized to more than one signal to correct for, as it does not rely on a joint modeling of all types of data at hand. Furthermore the segmentation used in the correction step enables us to deal with signals with different probe densities. Finally, this correction approach allowed us to keep all tumors in the study, as opposed to (Stransky et al., 2006) were tumors with CNV in a given region were excluded when analysing this region.

Also, prior information on genes or regions could be accounted for in the segmentation step. Indeed, the likelihood $\widehat{\mathcal{L}}(\tau, \tau')$ associated with a given region can be reweighted or penalized, the dynamic programming algorithm then applies with the same computational complexity.

Acknowledgements

Appendix

A   Proof of Lemma (1)

Throughout this proof, we drop index k for sake of clarity. For a region with length ℓ, the covariance matrix Σ in Equation (1) can be rewritten as:

$$\Sigma = (1 - \rho)I + \rho J$$

where I stand for the $\ell \times \ell$ identity matrix and J for the $\ell \times \ell$ matrix with all entries equal to one. The inverse of this matrix has the form $\Sigma^{-1} = aI + bJ$ where

$$a = \frac{1}{1-\rho}, \qquad b = -\frac{\rho}{(1-\rho)\left[1-\rho+\ell\rho\right]}.$$

The determinant of Σ is

$$|\Sigma| = (1-\rho)^{\ell-1}(1-\rho+\ell\rho)$$

and the trace term $\operatorname{tr}\left[Y\Sigma^{-1}(Y)^{\top}\right]$ in equation (4) yields

$$\operatorname{tr}\left(Y(aI + bJ)Y^{\top}\right) = an\ell + bn\sum_{j=1}^{\ell}\sum_{k=1}^{\ell}\hat{G}_{jk}.$$

Combining all the above gives the log-likelihood for this region:

$$
\begin{aligned}
-2\log\mathcal{L} &= n\left[\log\left(1-\rho+\ell\rho\right)+(\ell-1)\log\left(1-\rho\right)\right] \\
&+\frac{n\ell}{1-\rho}-\frac{\rho n\sum_{j}^{\ell}\sum^{\ell}\hat{G}_{jk}}{(1-\rho)\left[1-\rho+\ell\rho\right]}.
\end{aligned}
$$

Optimizing this function wrt ρ gives the formula of the MLE (7). Pluging this estimate into the same function gives the contrast function given in (8).

B   Distribution of the test statistic

Note $Y_i^{(k)} = (Y_{i,\tau_{k-1}+1}, ..., Y_{i,\tau_k})^T$. Using the same notations as in Section 3, one has

$$Y_i^{(k)} \sim \mathcal{N}\left(0, \sigma_k\right) \Rightarrow Y_{i\bullet}^{(k)} \sim \mathcal{N}\left(0, \frac{1+(p_k-1)\rho_k}{p_k}\right)$$

Because variables $Y_i^{(k)}$, $i = 1, ..., n$ are i.i.d so do variables $Y_{i\bullet}^{(k)}$, and consequently

$$\sum_i^n\left(Y_{i\bullet}^{(k)} - Y_{\bullet\bullet}^{(k)}\right)^2 \sim \frac{1+(p_k-1)\rho_k}{p_k}\chi_{n-1}^2$$

References

Aldred, P., E. Hollox, and J. Armour (2005). Copy number polymorphism and expression level variation of the human alpha-defensin genes defa1 and defa3. *Hum Mol Genet. 14*(14), 2045–52.

Anderson, T. (1958). *An introduction to multivariate statistical analysis* (1st ed.). Series in Probability and Statistics. Wiley.

Bickmore, W. (2013). The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet. 14*, 67–84.

Bien, J. and R. J. Tibshirani (2011). Sparse estimation of a covariance matrix. *Biometrika 98*(4), 807–820.

Clark, S. J. (2007). Action at a distance: epigenetic silencing of large chromosomal regions in carcinogenesis. *Human Molecular Genetics 16*(R1), R88–R95.

Cohen, B. A., R. D. Mitra, J. D. Hughes, and G. M. Church (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature genetics 26*(2), 183–186.

Coppe, A., G. A. Danieli, and S. Bortoluzzi (2006). REEF: searching regionally enriched features in genomes. *BMC Bioinformatics 7*(1), .

Dai, M., P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic acids research 33*(20), e175–e175.

De, S. and M. Babu (2010). Genomic neighbourhood and the regulation of gene expression. *Curr Opin Cell Biol. 22*(3), 326–33.

Dottorini, T., P. Palladino, N. Senin, T. Persampieri, R. Spaccapelo, and A. Crisanti (2013). CluGene: A bioinformatics framework for the identification of co-localized, co-expressed and co-regulated genes aimed at the investigation of transcriptional regulatory networks from high-throughput expression data. *PloS One 8*(6), .

Frigola, J., J. Song, C. Stirzaker, R. Hinshelwood, M. Peinado, and S. Clark (2006). Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nat Genet. 38*(5), 540–9.

Heidenblad, M., D. Lindgren, T. Jonson, F. Liedberg, S. Veerla, G. Chebil, S. Gudjonsson, Å. Borg, W. Månsson, and M. Höglund (2008). Tiling resolution array cgh and high density expression profiling of urothelial carcinomas delineate genomic amplicons and candidate target genes specific for advanced tumors. *BMC medical genomics 1*(1), 3.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*(2), 249–264.

Lai, W., M. Johnson, R. Kucherlapati, and P. J. Park (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics 21*(19), 3763–3770.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing 85*(8), 1501 – 1510.

Lavielle, M. and G. Teyssière (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal 46*(3), 287–306.

Leday, G. G., A. W. van der Vaart, W. N. van Wieringen, M. A. van de Wiel, et al. (2013). Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines. *The Annals of Applied Statistics 7*(2), 823–845.

Lemay, D. G., W. F. Martin, A. S. Hinrichs, M. Rijnkels, J. B. German, I. Korf, and K. S. Pollard (2012). G-NEST: a gene neighborhood scoring tool to identify co-conserved, co-expressed genes. *BMC Bioinformatics 13*(1), .

Levina, E., A. Rothman, J. Zhu, et al. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics 2*(1), 245–263.

Lu, C., J. Feng, Z. Lin, and S. Yan (2013). Correlation adaptive subspace segmentation by trace lasso. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 1345–1352. IEEE.

Menezes, R. X., M. Boetzer, M. Sieswerda, G.-J. B. van Ommen, and J. M. Boer (2009). Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics 10*(1), .

Network, C. G. A. R. et al. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature 507*(7492), 315–322.

Nilsson, B., M. Johansson, A. Heyden, S. Nelander, and T. Fioretos (2008). An improved method for detecting and delineating genomic regions with altered gene expression in cancer. *Genome Biol 9*(1), .

Picard, F., E. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and S. Robin (2011). Joint segmentation,calling, and normalization of multiple CGH profiles. *Biostatistics*, 1–16.

Picard, F., S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics 6*(27), 1.

Rebouissou, S., I. Bernard-Pierrot, A. de Reyniès, M.-L. Lepage, C. Krucker, E. Chapeaublanc, A. Hérault, A. Kamoun, A. Caillault, E. Letouzé, et al. (2014). Egfr as a potential therapeutic target for a subset of muscle-invasive bladder cancers presenting a basal-like phenotype. *Science translational medicine 6*(244), 244ra91–244ra91.

Reyal, F., N. Stransky, I. Bernard-Pierrot, A. Vincent-Salomon, Y. de Rycke, P. Elvin, A. Cassidy, A. Graham, C. Spraggon, Y. Désille, et al. (2005). Visualizing chromosomes as transcriptome correlation maps: evidence of chromosomal domains containing co-expressed genes: study of 130 invasive ductal breast carcinomas. *Cancer Research 65*(4), 1376–1383.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Månér, H. Massa, M. Walker, M. Chi, et al. (2004). Large-scale copy number polymorphism in the human genome. *Science 305*(5683), 525–528.

Seifert, M., K. Abou-El-Ardat, B. Friedrich, B. Klink, and A. Deutsch (2014). Autoregressive higher-order hidden markov models: Exploiting local chromosomal dependencies in the analysis of tumor expression profiles. *PloS One 9*(6), .

Spellman, P. T. and G. M. Rubin (2002). Evidence for large domains of similarly expressed genes in the drosophila genome. *Journal of Biology 1*(1), 5.

Sproul, D., N. Gilbert, and W. A. Bickmore (2005). The role of chromatin structure in regulating the expression of clustered genes. *Nature Reviews Genetics 6*(10), 775–781.

Stransky, N., C. Vallot, F. Reyal, I. Bernard-Pierrot, S. de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C. Spraggon, et al. (2006). Regional copy number–independent deregulation of transcription in cancer. *Nature genetics 38*(12), 1386–1396.

Teschendorff, A. E., F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck (2013). A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics 29*(2), 189–196.

Tibshirani, R. and P. Wang (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics 9*(1), 18–29.

Vallot, C., N. Stransky, I. Bernard-Pierrot, A. Hérault, J. Zucman-Rossi, E. Chapeaublanc, D. Vordos, A. Laplanche, S. Benhamou, T. Lebret, et al. (2011). A novel epigenetic phenotype associated with the most aggressive pathway of bladder tumor progression. *Journal of the National Cancer Institute 103*(1), 47–60.

van Wieringen, W. N., J. Berkhof, and M. A. van de Wiel (2010). A random coefficients model for regional co-expression associated with DNA copy number. *Statistical Applications in Genetics and Molecular Biology 9*(1).

Xiao, G., C. Reilly, and A. B. Khodursky (2009). Improved detection of differentially expressed genes through incorporation of gene locations. *Biometrics 65*(3), 805–814.

Yi, Y., J. Mirosevich, Y. Shyr, R. Matusik, and A. L. George Jr (2005). Coupled analysis of gene expression and chromosomal location. *Genomics 85*(3), 401–412.

Zhang, Q., L. Ding, D. E. Larson, D. C. Koboldt, M. D. McLellan, K. Chen, X. Shi, A. Kraja, E. R. Mardis, R. K. Wilson, et al. (2010). CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics 26*(4), 464–469.